Vo Declaration
Exhibit A

Exhibit 1

Sealed Version

Message

From: Iliyan Zarov [[redacted]meta.com]
Sent: 11/20/2023 8:18:37 PM
To: Iliyan Zarov[redacted]@meta.com]; Nikolay Bashlykov [redacted]meta.com]
Subject: Message summary [{"otherUserFbId":100083037918592,"threadFbId":null}]

Nikolay Bashlykov (11/20/2023 11:38:31 PST):
>hi Iliyan! Robert mentioned that you might have already the OpenWebMath dataset, could you point me to the path?
>https://arxiv.org/pdf/2310.06786.pdf

Iliyan Zarov (11/20/2023 11:44:27 PST):
>hi Nikolay, sure, it's here - /fsx-llm/[redacted]

Nikolay Bashlykov (11/20/2023 11:46:08 PST):
>@silent amaxing, thx! did you run any experiments on it already?

Iliyan Zarov (11/20/2023 11:47:25 PST):
>Yep, I'm using it for a training/finetuning run I'm doing for reasoning, gives a pretty decent boost on stuff like gsm8k and math

Nikolay Bashlykov (11/20/2023 11:48:38 PST):
>cool! do you have any documents to read on that?

Nikolay Bashlykov (11/20/2023 11:49:26 PST):
>also do you know any other good math/reasoning datasets I should look into? or actually any other non-llama-2 datasets that you think are valuable?

Iliyan Zarov (11/20/2023 11:54:48 PST):
>nothing detailed, I added it after I started training so I did some limited ablations to check that helps over and above libgen and the galactica datasets and it seems to do

Iliyan Zarov (11/20/2023 11:55:09 PST):
>well, libgen and galatica datasets are great but I guess you're fully aware of these ☺

Nikolay Bashlykov (11/20/2023 11:55:50 PST):
>yep ☺

Iliyan Zarov (11/20/2023 11:56:15 PST):
>I'll be trying adding some synthetic data as well sampled from the improved model, but it's still early days on that

Nikolay Bashlykov (11/20/2023 11:56:36 PST):
>that's interesting!

Nikolay Bashlykov (11/20/2023 11:57:10 PST):
>btw, is it the whole dataset? it's like 27G, but I thought 15B tokens should be ~60G

Iliyan Zarov (11/20/2023 12:01:27 PST):
>it should be the whole thing yeah - I got it off hugging face. They report 6.32M rows of data and that's how many there are in the files total

Iliyan Zarov (11/20/2023 12:02:29 PST):
>do you have any other math/reasoning datasets you've been considering?

Nikolay Bashlykov (11/20/2023 12:04:36 PST):
>not really, but if you haven't talked to Viktor Kerkez, you can ask him - probably he has smth on top of the galactica

Iliyan Zarov (11/20/2023 12:05:46 PST):
>ah yes thanks, I'm in touch with him, he's been refreshing/expanding on the galactica corpus

Nikolay Bashlykov (11/20/2023 12:08:11 PST):
>also I updated the libgen dataset - basically removed some of the artifacts and outliers. I still need to move it to S3, but it's already on RSC:
>
>fiction: /[redacted]ibgen/fiction/fiction_en_20231120
>
>scitech:[redacted]libgen/scitech/scitech_en_20231120
>
>scimag:[redacted]libgen/scimag/scimag_20231120

Iliyan Zarov (11/20/2023 12:10:02 PST):
>oh great! what did you remove? I took the scitech version from you and I think I only removed some of the last pages with some heuristics as there were lots of indices, references, etc

Nikolay Bashlykov (11/20/2023 12:15:18 PST):
>basically some bad docs (based on token distribution), repetitions, copyright, emails:
https://docs.google.com/document [redacted]

Iliyan Zarov (11/20/2023 12:16:46 PST):
>Awesome! I'll try out the modified version. Are you copying it to s3/AWS? If not I'll prob copy scitech myself tomorrow

Nikolay Bashlykov (11/20/2023 12:18:30 PST):
>yep, copying it now, will update you once done

Nikolay Bashlykov (11/20/2023 12:18:37 PST):
>to S3